Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps

William J. Murphy,[1] Guillaume Bourque,[2] Glenn Tesler,[3] Pavel Pevzner[3] and Stephen J. O'Brien[1]*

[1]Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702, USA
[2]Centre de Recherches Mathématiques, Université of Montréal, Montréal, H3C 3J7, Canada
[3]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093-114, USA
**Correspondence to*: Tel: +1 301 846 7478; Fax: +1 301 846 6327; E-mail: murphywi@mail.ncifcrf.gov

Date received (in revised form): 19th August 2003

Abstract

Rapidly developing comparative gene maps in selected mammal species are providing an opportunity to reconstruct the genomic architecture of mammalian ancestors and study rearrangements that transformed this ancestral genome into existing mammalian genomes. Here, the recently developed Multiple Genome Rearrangement (MGR) algorithm is applied to human, mouse, cat and cattle comparative maps (with 311–470 shared markers) to impute the ancestral mammalian genome. Reconstructed ancestors consist of 70–100 conserved segments shared across the genomes that have been exchanged by rearrangement events along the ordinal lineages leading to modern species genomes. Genomic distances between species, dominated by inversions (reversals) and translocations, are presented in a first multispecies attempt using ordered mapping data to reconstruct the evolutionary exchanges that preceded modern placental mammal genomes.

Keywords: *genome evolution, synteny, mammals, ancestral genome*

Introduction

Great strides in understanding the evolutionary history of whole vertebrate genomes have been made over the past decade with the explosion of comparative mapping and sequencing data from diverse organisms.[1–7] Comparative maps from birds and mammals, coupled with recent human and mouse genomic sequences, have already provided many interesting insights into the evolutionary patterns and potential forces behind chromosomal rearrangements in vertebrates.[5–9] Previous vertebrate gene order comparisons have been limited to single chromosome comparisons of multiple genomes[5,6,10–12] or defining conserved segments between two whole genomes, however, rather than between multiple whole genomes.[3–6,11,13–16]

Comparative studies to identify and quantify the extent of conserved segments between two genomes are often based on the breakpoint analysis approach pioneered by Nadeau and Taylor.[17] These early studies of rearrangements between human and mouse genomes considered breakpoints independently, without revealing combinatorial dependencies between related breakpoints. Kececioglu and Sankoff[18] were the first to explore the importance of dependencies between breakpoints, and developed an approximation algorithm for the reversal distance problem (eg studies of rearrangements in unichromosomal genomes). Hannenhalli and Pevzner[19,20] developed a polynomial-time algorithm for the reversal distance problem, which was extended to the genomic distance problem of finding a most parsimonious scenario for multichromosomal genomes under inversions (reversals), translocations, fusions and fissions of chromosomes.[21–23]

Although these studies provided efficient algorithms to study rearrangements between two genomes, integrating data from multiple genomes (genome phylogeny) poses a more difficult problem. Previous genome phylogeny analyses were based on breakpoint distances that measure the number of breakpoints between two genomes.[24–26] Bourque and Pevzner[27] proposed a new approach, the Multiple Genome Rearrangement (MGR) algorithm, based upon the reversal/genomic distance. The MGR applications demonstrated important advantages of the reversal/genomic distance over the breakpoint distance. One strength of this new method is that it is directly adaptable to multichromosomal genomes, a variable unexplored in breakpoint distance approaches to date. The method is applicable to any group of multichromosomal organisms with comparative mapping data on the same set of markers, and can provide an estimate of original synteny (an ancestral genome) in the organisms under study.[28,29] Recently, other methods

studying rearrangement scenarios using the reversal distance were developed[30,31] but, so far, these methods are restricted to the median problem of unichromosomal genomes.

Here, an expanded set of homologous syntenic markers between the human, cat and mouse genomes is analysed, along with a set shared between human, cat and cattle genomes. Moreover, we derive a parsimonious genome rearrangement scenarios for these species and the hypothetical ancestral genomes for these index species imputed. A comparison of these inferences with reconstructions of the ancestral placental mammal karyotype based on comparative cytogenetic approaches[8,32,33] were largely concordant, validating the MGR approach[27] for using moderately dense comparative maps across mammalian orders to define the exchanges that led to modern genome reorganisation in each lineage.

Supporting information on the two datasets (human–cat–cow and human–cat–mouse) has been posted at www.ingenta.com

Methods

MGR algorithm

The MGR algorithm developed by Bourque and Pevzner[27] reconstructs a rearrangement-based evolutionary tree, considering reversals (more commonly called inversions), translocations, fusions and fissions. MGR is based on the Hannenhalli–Pevzner theory[34] and a fast implementation of the multichromosomal genome rearrangement algorithm[22,23] called *GRIMM*. The MGR algorithm works in two stages. Assume one wishes to attempt to reconstruct the rearrangement scenario of m genomes. In the first stage, rearrangement events in genome i ($1 \leq i \leq m$), that bring it closer to each of the remaining $m - 1$ genomes, are iteratively carried out in a carefully selected order. The rearrangements performed in the first stage are very reliable.[27] In fact, when there are only three genomes ($m = 3$), all three genomes are converted into the real ancestor if the tree is additive. In the case of non-additive trees, the first step stops before converging to an ancestor and an intermediate genome, or preancestor, is left. Because the moves made to reach the preancestors in the first stage were made with the highest confidence, it can be argued that studying them can provide insights into the global rearrangement scenario. In the second stage, the conditions for rearrangements to be carried out are relaxed by choosing a rearrangement in genome i that brings it closer to $t = m - 2$ out of $m - 1$ other genomes. We stop once again if all genomes converge to an ancestor. Otherwise, the parameter t is further lowered. For a full description of the algorithm, see Bourque and Pevzner.[27]

In the context of genome rearrangements, genomes are typically viewed as signed permutations, where each integer corresponds to a unique gene/marker and the sign corresponds to its orientation. By contrast, comparative maps usually correspond to unsigned permutations — ie no information on the sign of the markers is available. Since no efficient algorithms for rearrangement analysis of unsigned permutations are available, Bourque and Pevzner[27] searched for strips in the unsigned permutations to infer the signed permutations from the original data.[35] A strip is two or more markers that appear consecutively in all genomes in the exact same order, or reversed order (to which we assign reversed signs), without any interruption by other markers. A marker that is not part of any strip is called a singleton and is dropped from the signed permutation due to uncertainty in its sign. Below, we propose a new, more flexible, method to recover the signed permutation from the comparative mapping data that uses clusters (two or more markers located closely to each other in all genomes) instead of strips. This new method is less sensitive to local mapping errors and to micro-rearrangements that can complicate the recovery of the global rearrangement scenario.

GRIMM-synteny algorithm for cluster generation

A particularly confounding variable in comparative genome analysis is the distinction between small micro-rearrangements that interrupt conserved segments and exceptional singleton markers that reflect imprecise map orders or mapping/assembly errors. Making this aspect more perplexing are recent comparisons of full genome sequences for mouse and human which show significantly more rearrangements than previously predicted, due to evidence of multiple micro-rearrangements within previously defined conserved segments.[3–5,36] Here, the notion of conserved segments is relaxed and the notion of a gene (marker) cluster introduced. Every cluster (comparable to a synteny block) corresponds to a set of markers located close to each other in each of the genomes under study. The order of markers within the cluster may vary from one genome to another, and may reflect mapping imprecision or actual micro-rearrangements.[37] Thus, clusters are the fragments of the genome that can be converted into conserved segments by micro-rearrangements (eg by reversals spanning relatively few markers). Local errors in comparative maps and micro-rearrangements make it non-trivial to find clusters.[25,38–40] Here, we describe the clustering algorithm using three genomes (human, cat and mouse) with comparative mapping data, but the algorithm applies to two or more genomes.[27,36]

To perform the multispecies genome comparisons, we first concatenate chromosomes in human, cat and mouse to form a single coordinate system for each genome based on n markers. The markers in each concatenation are assigned coordinates $1,2,\ldots,n$. A marker located at position h in human, c in cat and m in mouse is assigned a coordinate (h, c, m) that can be viewed as an element of a three-dimensional n by n by n grid. Triplets of chromosomes divide this grid into boxes

Table 1. Conserved markers, clusters and reversal distances computed with GRIMM-synteny and Multiple Genome Rearrangement Algorithm analysis of comparative gene maps of 470 Type I gene homologues aligned between human (H), mouse (M) and cat (C) genomes. The common ancestor of all three genomes is denoted A, while preancestors for human, mouse and cat are denoted H*, M* and C*, respectively. The total distance between the three genomes, d(H, M, C), is defined as d(H, M) + d(M, C) + d(C, H). The tree score is defined as d(A,H) + d(A,M) + d(A,C)

Distance threshold, *G*	**4**	**5**	**6**	**8**	**20**
No. of markers retained	276	345	379	409	432
% of markers used	59	73	81	87	92
No. of clusters	112	114	106	94	76
d(H, M, C)	222	234	216	201	160
d(A, H*) + d(H*, H)	19 + 10 = 29	18 + 9 = 27	19 + 10 = 29	15 + 9 = 24	11 + 6 = 17
d(A, C*) + d(C*, C)	13 + 15 = 28	13 + 11 = 24	13 + 8 = 21	10 + 12 = 22	13 + 5 = 18
d(A, M*) + d(M*, M)	25 + 41 = 66	31 + 49 = 80	32 + 40 = 72	21 + 43 = 64	24 + 32 = 56
Tree score	123	131	122	110	91

(the human, cat and mouse comparison has $24 \times 20 \times 21$ boxes). Every marker is on a triplet of chromosomes (one from human, one from cat and one from mouse). The distance between two points (h_1, c_1, m_1) and (h_2, c_2, m_2) from the same chromosome triplet (the same box) is the Manhattan distance $|h_2 - h_1| + |c_2 - c_1| + |m_2 - m_1|$. The distance between points from different chromosome triplets is defined as infinity.

MGR can be directly applied to all genetic markers shared by human, cat and mouse to find a rearrangement scenario; however, this scenario is likely to be flawed, since comparative maps will have some unreliably positioned markers that impute a false rearrangement. Therefore, we apply the GRIMM- synteny algorithm to filter out spurious markers that occur as isolated points (or 'small clusters') in a marker matrix. The GRIMM-synteny algorithm for comparative data invokes a distance threshold, *G*, as a parameter. The distance threshold is defined as the number of chromosomal interruptions below which markers are deemed to be part of the same synteny block.

GRIMM-synteny algorithm

(1) Form a marker graph whose vertex set is the set of markers.
(2) Connect vertices in the marker graph by an edge if the distance between them is smaller than the distance threshold *G*.
(3) Define clusters as connected components in the marker graph.
(4) Delete singletons (clusters with just one marker).
(5) Determine the cluster order and signs (orientation) for each genome.

We define the span of a cluster in human (or cat or mouse) as the interval between its minimum and maximum coordinates. Note that, although different clusters are not supposed to overlap in three dimensions, they often overlap in one dimension (ie their span intervals may overlap in human or cat or mouse). Therefore, defining the cluster order for intermingled clusters should be carried out with caution. To do this, we compute the centre of mass of all markers forming the cluster, and order clusters in human by the coordinates of their centres of mass. Cluster numbers are assigned according to their order on the human genome and then ordered in the other genomes in terms of these labels. We define rearrangements of markers within a cluster as micro-rearrangements, while rearrangements of the order and orientation of clusters are called macro-rearrangements.

Maximum distance threshold. We illustrate the influence of the maximum distance threshold G on the set of derived clusters in the case of three genomes A, B and C. Consider two markers, x and y, that are adjacent in all three genomes, either as x, y or as y, x. Their distance is $d(x, y) = 1 + 1 + 1 = 3$, and they will be placed in the same cluster only if $G \geq 4$. Conversely, distances larger than 3 indicate that a pair of markers fails to be adjacent in one or more genomes. Hence, the threshold, G, limits the maximum number of chromosomal interruptions $d(x, y)$, between markers x and y across m genome comparisons.

Recall that a strip is a sequence of markers $x, y, \ldots, z$ that appear consecutively or reversed in all three genomes, without interruption by other markers. At $G < 4$, each marker forms its own singleton cluster and is deleted. At $G = 4$, each strip forms its own cluster. As G increases, some clusters may be merged together to form a larger cluster with micro-rearrangements. An example of this is shown in Figure 1.

Thus, for m genomes, $G \leq m$ puts each marker into its own singleton cluster that is deleted. $G = m + 1$ puts each

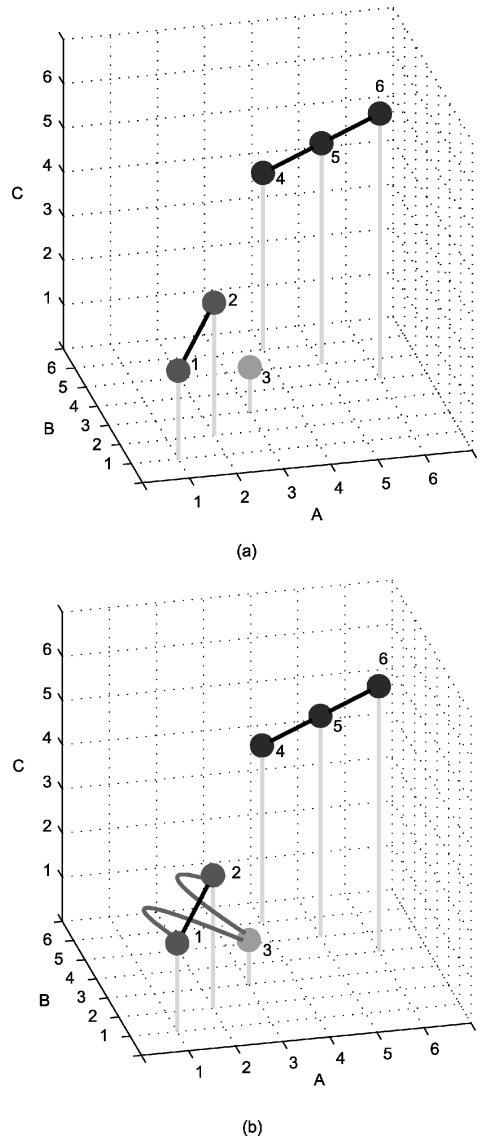

Figure 1. Illustrating the effect of the distance threshold, G, on cluster formation. Suppose genome A has marker order 1,2,3,4,5,6; genome B has 1,2,3,6,5,4; and genome C has 3,1,2,4,5,6. The strips are [1,2], [3], [4,5,6]. The clusters at $G = 4$ (a) are [1,2] and [4,5,6] (the singleton [3] is deleted). At $G = 5$ (not shown), some of these are combined together. Specifically, $d(2,3) = 1 + 1 + 2 = 4 < 5$, so an edge is added between markers 2 and 3, joining their clusters together. The clusters at $G = 5$ are [1,2,3] and [4,5,6] and the order within the clusters varies by genome, giving micro-rearrangements. At $G = 6$ and 7 (b), edges are added within clusters, but not between clusters, so clusters do not change. At $G = 8$ (not shown), two edges are added that would join the clusters into [1,2,3,4,5,6] Specifically, $d(2,4) = 2 + 4 + 1 = 7 < 8$ and $d(3,4) = 1 + 3 + 3 = 7 < 8$.

strip into its own cluster. $G = m + 2$ allows for clusters that form a strip in all but one genome, which instead has a pair of adjacent markers in that strip which are inverted (there can be multiple inverted pairs within a cluster, as long as no two pairs are adjacent). Therefore, increasing the value of G allows for clusters with more complex micro-rearrangements.

Comparative mapping data

Feline–human comparative mapping data (590 shared coding gene markers) have been described by Murphy *et al.* and Menotti-Raymond *et al.*[11,41] Human–mouse comparative mapping data were derived online, from http://www.ncbi.nlm.nih.gov/Homology. Cattle–human comparative mapping data were derived from Band *et al.*[15]and associated mouse data were derived from the previously listed mouse databases. For cases where mapped homologous loci did not exist for a given species pair, we found the most physically proximal human gene, which was taken as a 'virtual' coordinate to find a mapped mouse homologue in genetic or radiation hybrid (RH) maps. Cattle homologues were considered equivalent 'common' markers if their human homologue resided within 20 centirays (map units) of the human–cat anchors and were consistent with previously defined blocks of human–cattle synteny.[15] In a few cases, the virtual marker was extended to 50 centirays, but only where it was certain that there were no violations of previously defined syntenies. For this analysis, we assembled two comparative datasets:

(1) Human–mouse–cat (470 shared markers), which represented two conserved (few rearrangements from the ancestral placental genome[8]) mammalian genomes (human and cat) with one significantly reshuffled mammalian genome (mouse).
(2) Human–cat–cow (311 shared markers), which represented two conserved mammalian genomes (human and cat) and one moderately reshuffled mammalian genome (cow).[8]

The number of identified homologous mapped markers (actual plus virtual) between the species pairs human–cat, human–cow and cat–mouse and cat–cow, was 551, 633, 470 and 311, respectively.

Results

Human–mouse–cat dataset

The genomic maps of homologous markers were first compared between human, mouse and cat using the MGR and GRIMM-synteny algorithms. The comparison involved 470 Type 1 coding gene markers with MGR distance threshold parameters set at $G = 4, 5, 6, 8$ and 20 (Table 1). The results reveal several important patterns that can be interpreted in a comparative genomics context. First, increasing the distance threshold typically results in an increase in the number of

Table 2. Comparison of the Multiple Genome Rearrangement (MGR) algorithm-derived syntenies found in the common ancestors of the human–cat–mouse (HCM) and the human–cat–cow (HCC) datasets, with predicted syntenies based on comparative cytogenetic analyses (left-hand column[8]). MGR analyses were performed using the indicated distance threshold, G

Predicted syntenies[8]	HCM *G* = 4	HCM *G* = 5	HCM *G* = 6	HCC *G* = 4	HCC *G* = 5	HCC *G* = 6
3 & 21	+,f	+	+	+	+	+
4 & 8p	+	+	+	n.c.a.	–	–
7a/16p	n.c.a.	n.c.a.	+	+	+	+
12 & 22a	+,f	+,f	+,f	–	–	–
12 & 22b	–	–	n.c.a.	n.c.a.	n.c.a.	n.c.a.
14 & 15	+,f	+	–	+	–	–
16q/19q	–	+	+	–	+	+
1p	+,f	+	+,f	+	+	+
1q	–	–	–	–	–	–
2p	+,f	+,f	+,f	+,f	+,f	+,f
2q	+	+	+	+	+	+,f
5	+,f	+,f	+,f	+,f	+	–
6	+	+	+	+	+	+
7b	+,f	–,f	+,f	+,f	+,f	+,f
8q	+	+	+	+	+	+
9	+	+	+	+	+	+
10p	+,f	+,f	+,f	+,f	+,f	+,f
10q	+,f	–	+,f	+,f	+,f	+,f
11	–	–	+,f	+	+	+
13	+,f	+,f	+,f	+	+	+,f
17	+	+	+	+	+	+
18	+	+,f	+,f	+	+	+
19p	+,f	+,f	+,f	+,f	+	+
20	+,f	+,f	+,f	+,f	+,f	+
X/f	+	+	+	+	+	+

' + ' means synteny is intact in the ancestor, ' – ' means synteny is disrupted in the ancestor, ' + ,f ' means synteny is intact and fused to another chromosome in the ancestor. n.c.a. = no chromosome available, due to lack of shared markers defining that conserved segment.

markers returned in clusters, as fewer singletons are dropped. Another consequence of the threshold increase is that the number of clusters typically decreases, as does the overall rearrangement distance. This is the result of reducing the number of local rearrangements due to poor mapping resolution of tightly linked markers, or derived micro-rearrangements, in the mouse genome. At very high thresholds (eg $G = 20$), almost all internal inversions are not counted, in many cases collapsing entire chromosomes into single conserved segments. We show results at high thresholds only to demonstrate the failure to resolve chromosome associations (see below) with a few diagnostic markers, while enhancing recovery of single chromosome syntenies (Table 2). In practice, however, we do not advocate

Table 3. Conserved markers, clusters and reversal distances computed with GRIMM-synteny and the Multiple Genome Rearrangement Algorithm analysis of comparative gene maps of 311 Type I gene homologues aligned between human (H), cat (Ct) and cow (Cw) genomes. The common ancestor of all three genomes is denoted A, while preancestors for human, cat and cow genomes are denoted H*, Ct* and Cw*, respectively. The total distance between the three genomes, d(H, Ct, Cw), is defined as d(H, Ct) + d(Ct, Cw) + d(Cw, H). The tree score is defined as d(A, H) + d(A, Ct) + d(A, Cw)

Distance threshold, *G*	**4**	**5**	**6**	**8**	**20**
No. of markers used	248	262	276	286	298
% of markers used	80	84	89	92	96
No. of clusters	81	74	70	60	44
d(H, Ct, Cw)	129	126	119	98	63
d(A, H*) + d(H*, H)	4 + 10 = 14	3 + 11 = 14	4 + 12 = 16	1 + 12 = 13	2 + 6 = 8
d(A, Ct*) + d(Ct*, Ct)	8 + 14 = 22	8 + 17 = 25	6 + 15 = 21	9 + 11 = 20	3 + 7 = 10
d(A, Cw*) + d(Cw*,Cw)	12 + 22 = 34	11 + 18 = 29	10 + 17 = 27	7 + 13 = 20	5 + 11 = 16
Tree score	70	68	64	53	34

using such high thresholds, as they result in loss of almost all intrachromosomal detail. A chromosome association is defined as clusters of two different human chromosomes that are adjacent on a single chromosome in another genome (ie fragments of human chromosomes 14 and 15 fused together (denoted 14/15) on cat chromosome B3) or in an ancestor.

Table 2 illustrates the sensitivity of the algorithm to threshold in recovery of ancestral chromosomes predicted by previous studies on chromosome painting and comparative mapping data.[8] It should be noted that previous studies were based on lower-resolution datasets generated for much larger sets of mammalian species (20–40 species from as many as eight placental orders). In general, increasing the threshold, *G*, tends to improve the consistency of the overall reconstruction with previous chromosomal syntenies.

Figure 2 depicts a reconstructed ancestral genome from which the human, cat and mouse genomes descended, based on MGR-GRIMM ($G = 6$). The putative three-species ancestor contains 19 autosomes and the sex chromosomes, and shares a number of chromosomes and chromosome associations hypothesised to be in the ancestral placental mammal: these include associations 3/21 (human chromosome 3 fused to human chromosome 21), 4/8p, 7/16p, 16q/19q and single chromosome syntenies 2q, 8q, 9 and 17. This reconstruction differs from previous hypotheses by lacking, for example, the 14/15 chromosome association and one of the two 12/22 associations. If, however, the three preancestors (defined here as genomes on the path towards the ancestor on which rearrangements have only been performed with the highest confidence) are examined at threshold 4, there is evidence of these predicted associations in at least one of the preancestors (see supporting information at www.ingenta.com).

Human–cat–cow dataset

Table 3 shows the results of applying GRIMM-synteny and MGR to the 311 marker human–cat–cow dataset. As observed with the previous dataset, increasing the thresholds tends to add more markers but decreases conserved segment resolution. This dataset also recovers most of the human chromosome associations predicted in the placental ancestor, although fewer markers resulted in loss of some of the segments within the 4/8p and 12/22 associations (Table 2 and Figure 3). By contrast with the human–mouse–cat dataset, the more conserved human–cat–cow genome triple, with lower and more equal distances (Table 3), recovers more of the single human chromosome syntenies at lower thresholds (eg 4 and 5), while threshold 6 shows more of these single syntenies instead as associations (eg 13 with 5 and 2p + q). All datasets, descriptions of clusters and results from analyses of both human–cat–mouse and human–cat–cow datasets can be found in the supporting information at www.ingenta.com

Proportion of the various types of rearrangements

Table 4 shows a comparison of the proportions of each type of rearrangement at the varying thresholds for the human–mouse–cat versus the human–cat–cow datasets. Reversals (inversions) represent a very frequent category of rearrangement event in both datasets. The fact that this event category is even more common in the human–cat–cow dataset than in the human–cat–mouse dataset is consistent with previous analyses of mammalian comparative maps.[28] Recent human–mouse genomic sequence comparisons, however,[3–5] reveal that intrachromosomal rearrangements (reversals) are the most frequent rearrangement event, as will probably become more evident in the human–cat–mouse rearrangement scenario as the number

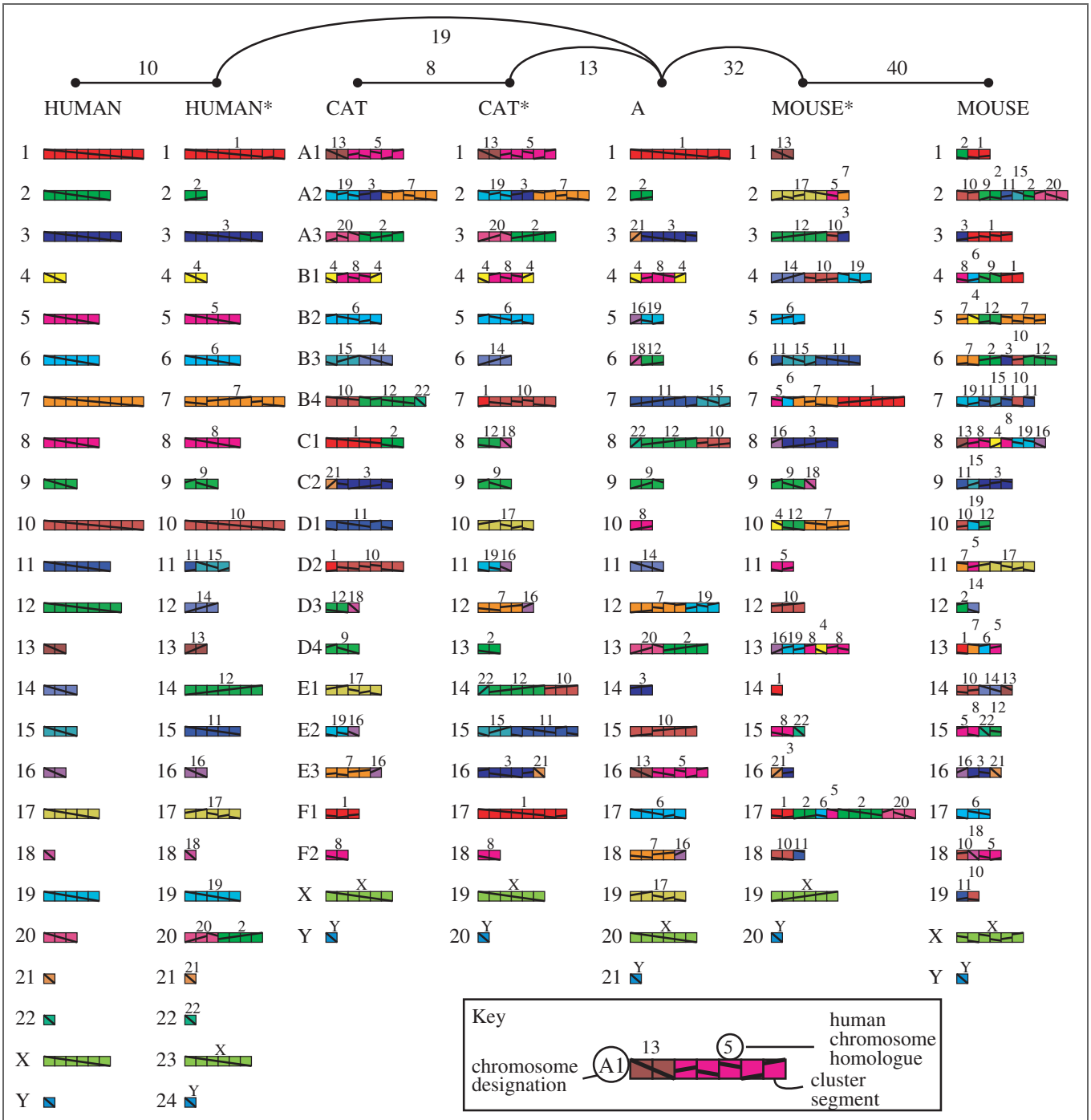

Figure 2. Chromosome syntenic organisation imputed by the Multiple Genome Rearrangement Algorithm for preancestors (denoted with asterisks) of human, of cat and of mouse, and for the reconstructed common ancestor (A) of all three starting genomes (human, cat and mouse). The data consisted of 379 common markers grouped into 106 clusters shared across the three starting genomes (human, cat and mouse) derived using a distance threshold of $G = 6$. The length of each chromosome is proportional to its number of cluster segments. Each human chromosome (and its component cluster segment [boxed]) is assigned a unique colour. Each cluster segment is traversed by a diagonal line (top left to bottom right) to indicate relative order and orientation within the blocks. The number above each coloured block refers to the corresponding human chromosome homologue. Species chromosome designations are shown to the left. At the top of the figure, the phylogram indicates the number of rearrangements required to convert one genome into the other.

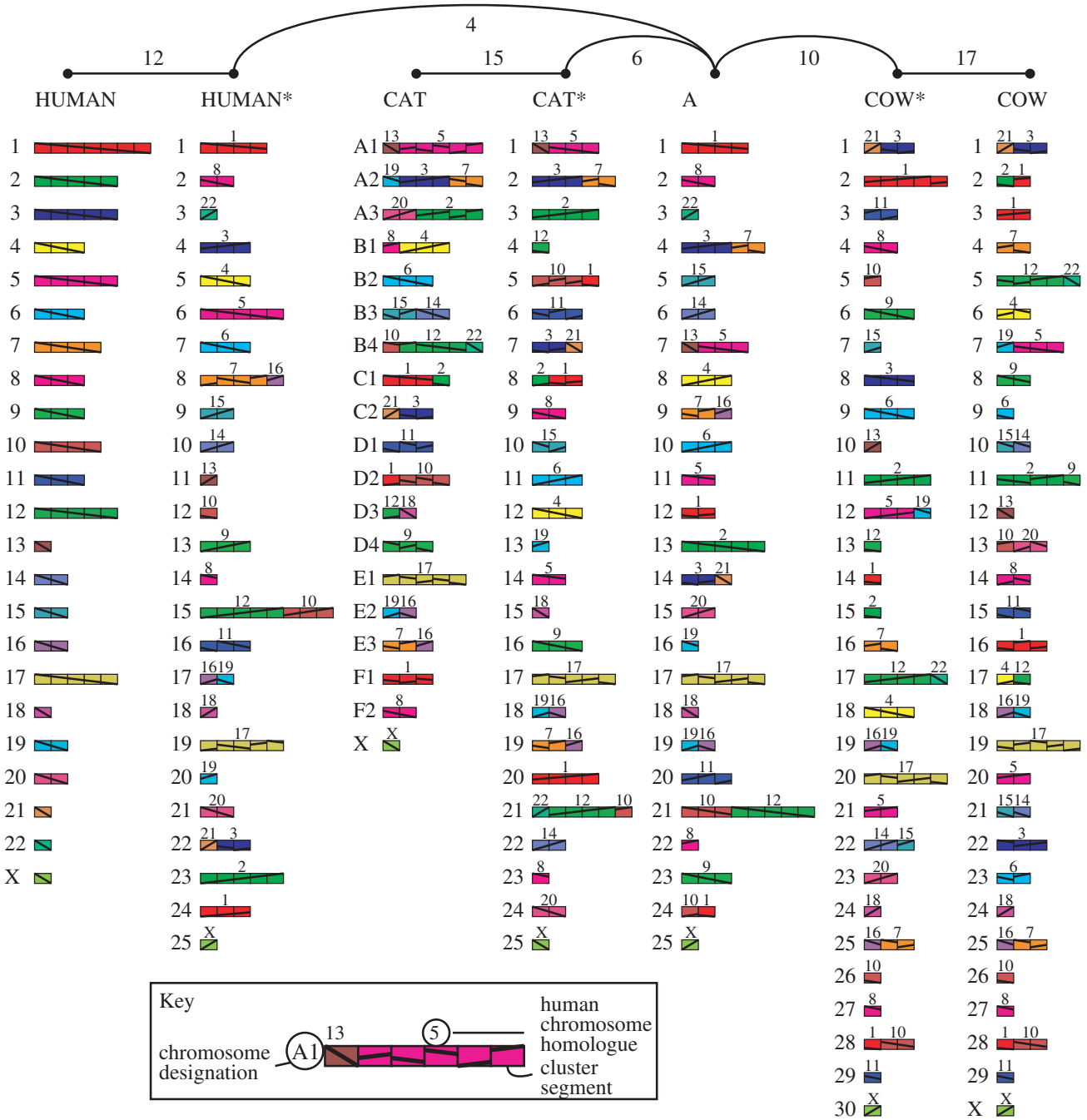

Figure 3. Chromosome syntenic organisation imputed by the Multiple Genome Rearrangement Algorithm for preancestors (denoted with asterisks) of human, of cat and of cow, and for the reconstructed common ancestor (A) of all three starting genomes (human, cat and cow). The data consisted of 276 common markers grouped into 70 clusters shared across the three starting genomes (human, cat and cow) derived using a distance threshold of $G = 6$. The length of each chromosome is proportional to its number of cluster segments. Each human chromosome (and its component cluster segments [boxed]) is assigned a unique colour. Each cluster segment is traversed by a diagonal line (top left to bottom right) to indicate relative order and orientation within the blocks. The number above each coloured block refers to the corresponding human chromosome homologue. Species chromosome designations are shown to the left. On the top of the figure, the phylogram indicates the number of rearrangements required to convert one genome into the other.

Table 4. Proportion of different types of rearrangements for the human–cat–mouse and the human–cat–cow datasets

	Distance threshold, *G*	4	5	6	8	20
Human–cat–mouse	% reversals	38.2	38.2	36.9	23.6	7.7
	% translocations	57.7	58.8	59.0	70.9	87.9
	% fusions	1.6	2.3	2.5	0.9	3.3
	% fissions	2.4	0.8	1.6	4.5	1.1
Human–cat–cow	% reversals	45.7	42.6	34.4	26.4	5.9
	% translocations	40.0	38.2	45.3	49.1	55.9
	% fusions	8.6	7.4	7.8	9.4	14.7
	% fissions	5.7	11.8	12.5	15.1	23.5

of shared markers increases. As might be expected, increasing the threshold reduces the breakpoint distance by reducing the proportion of reversals. The proportion of fusions and fissions over all types of rearrangements is about 5 per cent for the human–mouse–cat dataset, but varies from 14.3 per cent to 38.2 per cent for the human–cat–cow dataset. The proportion increases in the second dataset because, while the overall distance is being reduced, the number of fusions and fissions cannot drop below a certain constant required to explain the varying number of chromosomes between the three species' genomes. Regardless of this, the proportions of fusions and fissions remain within the range for which MGR has been tested and performs well.[27]

Discussion

Using multispecies mammalian comparative maps, coupled with new computational tools for multichromosomal rearrangement analysis, we have been able to demonstrate the promise of generating ancestral chromosome architectures from small numbers of taxa and fewer than 500 shared markers. Our results using two three-taxa datasets (human–cat–mouse and human–cat–cow) reconstruct, under different assumptions about treating local mapping errors and micro-rearrangements, mammalian ancestral genomes containing most of the chromosome associations and syntenies hypothesised based on chromosome painting inferences.[8,32,33] Of course, if the number of species is increased, markers will improve upon the accuracy of the ancestor reconstructions and rearrangement scenarios.

Despite having fewer common markers, the human–cat–cow dataset recovers the single chromosome syntenies (eg 5 and 13) at a higher frequency than the human–mouse–cat dataset, where they tend to be intact yet fused to other chromosomes (Table 2 and Figure 3). This is best explained by the overall slower rate of change among these three species (Table 3) and the tendency of most of these chromosomes to be fused to other human syntenic regions in the rearranged mouse genome. This confirms the conclusion that increasingly additive trees produce more reliable ancestors[27] and suggests that inclusion of more slowly evolving genomes will aid in the reconstruction of placental ancestral genomes.

One finding of interest is that, even though the mouse is highly rearranged compared with most species, increasing the threshold of considered micro-rearrangements (which have occurred largely on the mouse lineage) allows the algorithm to compensate and converge on a relatively unshuffled ancestor. Although there are some unexpected ancestral chromosomes in different analyses of the human–cat–mouse dataset, most of these represent fusions of intact human chromosomes that are thought to have been distinct chromosomes in the placental ancestor. One example is the fusion of human 2p and 20 into a single ancestral chromosome in almost all analyses within and between both datasets. This 2p/20 association is found intact in the cat genome and is believed to be ancestral for carnivores.[8,42,43] This has never been found in another placental karyotype examined with molecular methods, except in mouse, where human 20 is syntenic with a small fragment of human chromosome 2p. In rare cases like this, the apparently common carnivore-rodent association is best explained by convergence through the extensive chromosomal scrambling observed in the mouse genome.[1,4] This is supported by inspection of the rat genome,[1,14] which does not share this association. As with any phylogenetic analysis, increasing taxon (genome) sampling will decrease the effects of homoplasy and increase the reliability of the tree and ancestral reconstruction.

Because MGR inferences are parsimony-based, saturation and long-branch attraction issues remain outstanding problems that will need to be addressed in future applications of this method to infer mammalian genome rearrangements. Therefore, the choice of genomes will affect chromosomal reconstructions, hence caution must be exercised when making interpretations from ancestors imputed with combinations of

slowly and rapidly evolving genomes. A good illustration of this principle is observed in the difficulty of recovering the 14/15 association with the human–mouse–cat dataset. Human chromosomes 14 and 15 are syntenic in the large majority of placental mammal genomes examined to date,[8,32,33] although this synteny has independently been lost in the human–ape lineage and the murid rodent lineage. Thus, two of three genomes in the human–mouse–cat dataset lack this chromosome association (otherwise widespread in mammals), resulting in difficulty in recovering this ancestral chromosome. It should be noted that the human–cat–cow dataset, where two of three genomes do have the 14/15 association, recovers this ancestral chromosome at low thresholds, although recovers it less well when the threshold is increased due to loss of marker resolution.

Increased marker density will ultimately improve reconstruction accuracy. This was suggested by the improvement of the current human–mouse–cat ancestor over a previously computed scenario using these same three species, but with a much smaller number of markers.[27] This result supports previous conclusions emphasising that the number of markers should exceed a certain threshold to provide reliable ancestral reconstructions.[27]

As the number of ordered comparative maps from different mammalian species increases, along with an increase in shared markers, it is expected that the reliability of the ancestral reconstructions (both whole chromosomes and orders within chromosomes) will be more accurate reflections of the ancestral mammalian genome. These advances will initially proceed from the mapping stage, where a broader taxonomic sampling from whole genome descriptions is currently available (or in development). The promise and application of this approach to multiple mammalian genomic sequences from several orders will surely provide the greatest accuracy and insight into whole genome evolution, as demonstrated by current human–mouse whole genome comparisons.[3,4,36]

References

1. The International Human Genome Sequencing Consortium (2001), 'Initial sequencing and analysis of the human genome', *Nature* Vol. 409, pp. 860–921.
2. Venter, J.C., Adams, M.D., Myers, E.W. *et al.* (2001), 'The sequence of the human genome', *Science* Vol. 291, pp. 1304–1351.
3. Mouse Genome Sequencing Consortium (2002), 'Initial sequencing and comparative analysis of the mouse genome', *Nature* Vol. 420, pp. 520–562.
4. Gregory, S.G., Sekhon, M., Schein, J. *et al.* (2002), 'A physical map of the mouse genome', *Nature* Vol. 418, pp. 743–750.
5. Mural, R.J., Adams, M.D., Myers, E.W. *et al.* (2002), 'A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome', *Science* Vol. 296, pp. 1661–1671.
6. Dehal, P., Predki, P., Olsen, A.S. *et al.* (2001), 'Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution', *Science* Vol. 293, pp. 104–111.
7. O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J. *et al.* (1999), 'The promise of comparitive genomics in mammals', *Science* Vol. 286, pp. 458–481.
8. Murphy, W.J., Stanyon, R. and O'Brien, S.J. (2001), 'Evolution of mammalian genome organization inferred through comparative gene mapping', *Genome Biol.* Vol. 2, pp. R00005–R00009.
9. Burt, D.W., Bruley, C., Dunn, I.C. *et al.* (1999), 'The dynamics of chromosome evolution in birds and mammals', *Nature* Vol. 402, pp. 411–413.
10. Yang, Y.P. and Womack, J.E. (1998), 'Parallel radiation hybrid mapping, a powerful tool for high-resolution genomic comparison', *Genome Res.* Vol. 8, pp. 731–736.
11. Murphy, W.J., Sun, S., Chen, Z.Q. *et al.* (2000), 'A radiation hybrid map of the cat genome: Implications for comparative mapping', *Genome Res.* Vol. 10, pp. 691–702.
12. Goldammer, T., Kata, S.R., Brunner, R.M. *et al.* (2002), 'A comparative radiation hybrid map of bovine chromosome 18 and homologous chromosomes in human and mice', *Proc. Natl Acad. Sci. USA* Vol. 99, pp. 2106–2111.
13. Schibler, L., Vaiman, D., Oustry, A. *et al.* (1998), 'Comparative gene mapping: A fine-scale survey of chromosome rearrangements between ruminants and humans', *Genome Res.* Vol. 8, pp. 901–915.
14. Watanabe, T.K., Bihoreau, M.T., McCarthy, L.C. *et al.* (1999), 'A radiation hybrid map of the rat genome containing 5,225 markers', *Nat. Genet.* Vol. 22, pp. 27–36.
15. Band, M.R., Larson, J.H., Rebeiz, M. *et al.* (2000), 'An ordered comparative map of the cattle and human genomes', *Genome Res.* Vol. 10, pp. 1359–1368.
16. Kumar, S., Gadagkar, S.R., Filipski, A. and Gu, X. (2001), 'Determination of the number of chromosomal segments between species', *Genetics* Vol. 157, pp. 1387–1395.
17. Nadeau, J.H. and Taylor, B.A. (1984), 'Lengths of chromosome segments conserved since divergence of man and mouse', *Proc. Natl Acad. Sci. USA* Vol. 81, pp. 814–818.
18. Kececioglu, J. and Sankoff, D. (1995), 'Exact and approximation algorithms for the inversion distance between two permutations', *Alogorithmica* Vol. 13, pp. 180–210.
19. Hannenhalli, S. and Pevzner, P. (1995), 'Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals)', in *Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing*, pp. 178–189.
20. Hannenhalli, S. and Pevzner, P. (1999), 'Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals)', *J. ACM* Vol. 46, pp. 1–27.
21. Hannenhalli, S., Pevzner, P. (1995), 'Transforming mice into men (polynomial algorithm for genomic distance problem)', in *Proceedings of the 36th IEEE Symposium on Foundations of Computer Science*, pp. 581–592.
22. Tesler, G. (2002), 'GRIMM: Genome rearrangements web server', *Bioinformatics* Vol. 18, pp. 492–493.
23. Tesler, G. (2002), 'Efficient algorithms for multichromosomal genome rearrangements', *J. Comp. Sys. Sci.* Vol. 65, pp. 587–609.
24. Blanchette, M., Bourque, G. and Sankoff, D. (1997), 'Breakpoint phylogenies', in Miyano, S. and Takagi, T., eds, *Genome Informatics Workshop*, University Academic Press, Tokyo, pp. 25–34.
25. Sankoff, D. and Blanchette, M. (1997), 'The median problem for breakpoints in comparative genomics', in *Lecture Notes in Computer Science*, Springer Verlag, New York, pp. 251–263.
26. Moret, B., Wyman, S., Bader, D. *et al.* (2001), 'A new implementation and detailed study of breakpoint analysis', in *Proceedings of the 6th Pacific Symphosium on Biocomputing*, pp. 583–594.
27. Bourque, G. and Pevzner, P. (2002), 'Genome scale evolution: Reconstructing gene orders in the ancestral species', *Genome Res.* Vol. 12, pp. 26–36.
28. Ehrlich, J., Sankoff, D. and Nadeau, J.H. (1997), 'Synteny conservation and chromosomal rearrangements during mammalian evolution', *Genetics* Vol. 147, pp. 289–296.
29. Ferretti, V., Nadeau, J.H. and Sankoff, D. (1996), 'Original synteny', in *Combinatorial Pattern Matching* (CPM'96), Vol. 1075 of

Lecture Notes in Comput. Sci. Springer Verlag, New-York, pp. 159–167.
30. Siepel, A.C., Moret, B.M.E. (2001), 'Finding an optimal inversion median: Experimental results', in *Proceedings of the First International Workshop on Algorithms in Bioinformatics (WABI, 2001)*, Vol. 2149 of *Lecture Notes in Comput. Sci.* Springer Verlag, New York, pp. 189–203.
31. Caprara, A. (2003), 'The reversal median problem', *INFORMS J. Comput.* Vol. 15, pp. 93–113.
32. Chowdhary, B.P., Raudsepp, T., Fronicke, L. and Scherthan, H. (1998), 'Emerging patterns of comparative genome organization in some mammalian species as revealed by Zoo-FISH', *Genome Res.* Vol. 8, pp. 577–589.
33. Wienberg, J., Froenicke, L. and Stanyon, R. (2000), 'Insights into mammalian genome organization and evolution by molecular cytogenetics', in Clark, M.S., ed., *Comparative Genomics*, Kluwer, Dordrecht, the Netherlands, pp. 207–244.
34. Pevzner, P. (2000), *Computational Molecular Biology: An Algorithmic Approach*, MIT Press, Cambridge, MA.
35. Pevzner, P., Tesler, G. (2003), 'Transforming men into mice: The Nadeau-Taylor chromosomal breakage model revisited', in *Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2003)*, pp. 247–256, Appendix B.
36. Pevzner, P.A. and Tesler, G. (2003), 'Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes', *Genome Res.* Vol. 13, pp. 37–45.
37. Sankoff, D., Ferretti, V. and Nadeau, J.H. (1997), 'Conserved segment identification', *J. Comput. Biol.* Vol. 4, pp. 559–565.
38. Fujibuchi, W., Ogata, H., Matsuda, H. and Kanehisa, M. (2000), 'Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping', *Nucleic Acids Res.* Vol. 28, pp. 4029–4036.
39. Lathe, W.C., Snel, B. and Bork, P. (2000), 'Gene context conservation of a higher order than operons', *Trends Biochem. Sci.* Vol. 25, pp. 474–479.
40. Rogozin, I.B., Makarova, K.S., Murvai, J. *et al.* (2002), 'Congruent evolution of different classes of non-coding DNA in prokaryotic genomes', *Nucleic Acids Res.* Vol. 30, pp. 2212–2223.
41. Menotti-Raymond, M., David, V.A., Chen, Z.Q. *et al.* (2003), 'Second-generation integrated genetic linkage/radiation hybrid maps of the domestic cat (*Felis catus*)', *J. Hered.*, Vol. 94, pp. 95–106.
42. Fronicke, L., Muller-Navia, J., Romanakis, K. and Scherthan, H. (1997), 'Chromosomal homeologies between human, harbor seal (*Phoca vitulina*) and the putative ancestral carnivore karyotype revealed by Zoo-FISH', *Chromosoma* Vol. 106, pp. 108–113.
43. Nash, W.G., Menninger, J.C., Wienberg, J. *et al.* (2001), 'The pattern of phylogenomic evolution of the Canidae', *Cytogenet. Cell Genet.* Vol. 95, pp. 210–224.